Philosophia (2015) 43:1067–1080
DOI 10.1007/s11406-015-9610-x

Michael Huemer's A Priori Defense of Metaethical Internalism

Sanford Levy[1]

Received: 17 March 2015 / Revised: 25 May 2015 / Accepted: 2 June 2015 /
Published online: 13 June 2015

Abstract Versions of internalism have played important roles in metaethics, for example, in defending irrealist options such as emotivism. However, internalism is itself as controversial as the views it is used to defend. Standard approaches to testing the view, such as thought experiments about amoralists, have failed to gain consensus. Michael Huemer offers a defense of internalism of a different kind which he calls the "argument from interpretation." He presents the argument as one Humeans could embrace, but versions could be accepted by others, including Huemer himself. The argument begins from the assumption that a certain principle of charity is true and knowable a priori. But it can only be known a priori if internalism is true. Hence internalism is true. In this paper I argue that this important argument fails. My main objection makes use of recent work in empirical psychology. Huemer needs the principle of charity to be known a priori. I argue that rather than being an a priori issue, it is an empirical one and that the empirical evidence is strong enough to undermine his argument for internalism.

Keywords Huemer · Internalism · Externalism · Principle of charity

There are a number of kinds of internalism discussed by metaethicists. Michael Huemer's version is that if one recognizes a reason of some kind for acting, then one must have at least some motivation to act in accordance with that reason (Huemer 2005). He refers to this as the "magnetism of values," but I will use the more common term "internalism." Externalists deny this and say that recognizing something as a reason for action does not necessarily motivate. Internalism plays a central role in metaethics especially in arguing for forms of metaethical irrealism. For example, Charles Stevenson (1937, 1963) used a version of internalism to defend emotivism and Gilbert Harman (1975) used a version to defend a kind of relativism. Huemer is an

✉ Sanford Levy
slevy@montana.edu

[1] Department of History and Philosophy, Montana State University, Bozeman, MT 59717, USA

intuitionist and a moral realist. It is often thought that views like his are incompatible with internalism since moral beliefs represent moral facts and there are no facts the recognition of which necessarily motivate. But I am not interested in how one might reconcile a realism like Huemer's with internalism. I am interested in his argument for internalism.

Several approaches to investigate whether or not internalism is true have been tried. Some people have thought internalism to be intuitively obvious, but many do not find it so. The issue has been investigated by thought experiments involving the conceptual possibility of amoralists and immoralists (e.g., Hare 1963 and Brink 1989). However, like the application of intuition to internalism itself, intuitions about amoralists and immoralists differ in what seem to be intractable ways. More recently empirical evidence about psychopaths has been brought to bear on the issue since the psychopath seems the closest thing in real life to an amoralist (e.g., by Prinz 2007). Again, this has not resolved the issue since the debate turns on difficult questions such as whether psychopaths really understand morality. Huemer offers a very different approach. As an intuitionist, he could claim that internalism is known to be true by intuition. But although he does appeal to intuition, it is not such a dogmatic appeal. Instead he appeals to a different intuition which forms the basis of what he calls the "argument from interpretation." Briefly, the argument begins from the assumption that a certain principle of charity is true and knowable a priori. But it can only be known a priori if internalism is true. Hence internalism is true. Although Huemer presents the argument as one he thinks his opponents, Humeans, should adopt, and he does not explicitly endorse it, it is one he could accept with a modification described below. Further, he defends premises of the argument with what he calls "plausible arguments." Accordingly, I will refer to the argument and its claims as Huemer's, again keeping in mind that he can only accept a modified version. In any event, the argument is important given the importance of principles of charity and given the failure of other approaches to resolve the dispute between internalists and externalists. In this paper I develop the argument more than Huemer does so as to have a clearer version to evaluate. I then argue that it fails. My main objection makes use of recent work in empirical psychology. For his argument from interpretation to work, Huemer needs his principle of charity to be a priori. I argue that rather than being an a priori issue, it is an empirical one and that the empirical evidence is enough to undermine his argument.[1]

Huemer's Internalism

I begin with Huemer's internalism. He says that "[t]he idea is that if a being can and does *recognize* a normative reason, then the being must have at least some tendency to be moved by that reason" (2005, 161). This requires explanation. First, Huemer often talks about our *having* a reason, but in the quotation just given, he speaks of *recognizing* a reason. That is the more accurate version. The principle requires motivation only when one recognizes a reason. It does not entail that we all recognize all reasons that exist or that reasons can motivate when not recognized. Second, it is not clear whether

[1] A number of authors have argued that various principles of charity are empirical. For example, see Stich (1985) Thagard and Nisbett (1983) and Henderson (1987).

recognizing a reason entails that one correctly recognizes it. "Recognize" can imply success, but it can also imply merely accepting something as a reason, which allows for error. Third, his internalism allows for some disconnect between one's recognized reasons for action and one's motives. He says that "[w]e typically have normative reasons and motivating reasons for the same things. . . But often the *strength* of our motives fails to match the strength of our normative reasons" (2005, 155–6). Fourth, Huemer's statement of internalism uses the word "must." There is a modal claim here. To make this clearer, I will state the principle as "Necessarily, whenever one has a normative reason for action, one has a motivating reason for action."

Huemer's Principle of Charity

The basic idea behind principles of charity is that we should try to interpret people's behavior in ways that makes those people rational. If on our interpretation they turn out irrational, or too irrational, that is good evidence that our interpretation is incorrect. Some of the earliest versions, including that of Quine, appear in discussions of translation, say, from unknown languages into our own language. For example, if our translations of someone's utterances attribute contradictory beliefs to that person, our translation is likely at fault. Other principles of charity have been proposed for purposes beyond translation, for example, for interpreting inferential practices. Huemer's argument for internalism turns on a principle of charity that has to do with the attribution of mental states. It asserts that "when you are trying to figure out what another person believes, desires, and so on, you look at his behavior and attribute to him the beliefs, desires, and so on, that would make that behavior make the most sense" (2005, 159). I assume this is shorthand for a more complex claim: we should attribute beliefs and desires that make the most sense in the light of a number of factors including behavior, perceptual capacities, past history, and so on. Perhaps that is part of what the unexplained "and so on" gestures towards. However, I will stick to Huemer's version for brevity.

Principles of charity come in strengths. The strongest say that we should never interpret people as irrational, or we should do so only with overwhelming evidence. More modest versions say that we should not interpret behavior as irrational without reasonably good evidence and, perhaps, only when we have an empirical account of what they are doing when they violate normative standards (Thagard and Nisbett 1983, 252). Huemer's principle is moderate. He writes, "[t]he principle [of charity] is not that all actions or beliefs are rational, but only that we should interpret others to be by and large rational, if possible. Sometimes people's statements and behavior provide evidence that they are irrational in certain respects" (2005, 160). This is sketchy. Clarification often comes from examples, but Huemer does not provide many in this section of his book. He says that when someone is eating a lot of chocolate, we should assume she is rational and so we should assume she knows she is eating chocolate and likes chocolate. More information might indicate, say, that she dislikes chocolate and only eats it because she believes it has health benefits. On the assumption that the person is irrational, we might hypothesize that she hates chocolate and is eating it specifically because she hates it. Here we have an irrational disconnect between her desires and her

actions. Or we might hypothesize that, while eating her chocolate, she suffers from the delusion that she is swimming the English Channel and that she likes swimming the Channel. The second and third interpretations explain her behavior but are ridiculous, short of serious evidence. I take it that the ridiculousness is an empirical matter. It is not likely that someone would make this sort of mistake and to do so would require a serious and unusual malfunction. It is analogous to the idea found in discussions of principles of charity in translation that errors about observation sentences are unlikely (Henderson 1987, 236). The idea, then, is that when attributing beliefs and desires to someone, we should assume an explanation of the first sort rather than an explanation of the second or third sorts absent good reason.

In sum, Huemer's principle of charity comes down to at least the following. In interpreting behavior in terms of beliefs, desires and so on, we should, in so far as we reasonably can, assume that people are not acting contrary to their desires and aversions and that they are not suffering from serious delusions. Some behavior will be explained by false beliefs if having them is required to make someone's belief-desire-behavior triple rational. Only if there is fairly significant evidence of irrationality, say of delusions or serious disconnects between desires and behavior, would we give this assumption up, though we do not know what that significant evidence involves. There are a number of interpretative issues about this principle, but we need not address them here.

Two Versions of the Principle of Charity

Earlier I said that Huemer presents the argument from interpretation for internalism as something a Humean should adopt, but that he himself is not a Humean. However, I also said he could accept it with a modification. This is important since many besides Huemer reject the Humean ideas in play here. The issue turns on what goes into a charitable interpretation of a person's mental states. In generalized form, the principle of charity says that we should assume, if possible, that a person is rational when interpreting her behavior. But there are a several versions possible depending on what one thinks can motivate behavior. Huemer's explanation of the argument from interpretation occurs in his discussion of the Humean view that motivation is rooted in desires and that belief alone cannot motivate. So the principle of charity is stated in Humean terms: we should attribute a belief and a desire to the agent that makes the most sense of her behavior, if possible. Huemer rejects the Humean view that desire is required for motivation. He identifies four motivators. There are appetites and emotional motivators which are linked to desires. But there are also prudential and impartial reasons for action. Huemer thinks that the belief that one has prudential or impartial reasons can motivate without being linked to desire. For example, desire and prudence can come apart when a student desires to party but decides to study because it is more prudent. Her belief that a prudential reason applies to her situation motivates without being linked to a desire. Given this, we can distinguish two versions of the principle of charity.

1. The Explicit Humean Version: We should attribute beliefs and desires to persons that make the most sense given their behavior.

2. The Implicit Huemerean Version: We should attribute beliefs, desires, prudential reasons and impartial reasons to persons that make the most sense given their behavior.

It is possible to construct an argument from interpretation using the implicit Huemerian version of the principle of charity analogous to the Humean version Huemer provides. But for simplicity, I stick to the Humean version since my objections to the argument are relevant to either version.

Huemer's Defense of Internalism

The argument begins with the claim that we know the principle of charity is true. If so, it can only be known empirically or a priori. But, Huemer says, it cannot be known empirically. For to know it empirically, we must determine whether or not people's beliefs and desires make sense in the context of their behavior. To do this, we need to know what their beliefs and desires are. But we cannot directly observe them. Nor can we infer them without relying on the principle of charity itself, which would be circular. So the assumption of rationality cannot be tested by observing people's behavior and is therefore known a priori.

The rest of the argument appears in a single passage. I will quote it in full and then reconstruct the argument.

> The principle of charity is almost universally accepted. But what is its basis? It seems that the principle could be correct only if we were justified in believing, in general, that people are by and large rational. But for the reasons just discussed, this belief cannot be justified by observation. So it must be a priori. How can that be? Here is a natural answer: because it is a necessary truth that whenever one has a normative reason for action, one has a motivating reason for action. Suppose the contrary: suppose, that is, that one can have a normative reason to do A but no motive for doing A. If this is possible, then it seems that it would also be possible for a person to lack motives, in general, for doing things he had normative reasons for doing. Such a person would be generally irrational. But if it is possible for a person to be generally irrational, then how could we know a priori that people are in fact by and large rational? (2005, 160–1)

This argument involves two main claims. The first is that not only should we interpret people as by and large rational, but they are by and large rational. The second is the principle of internalism itself.

1. The principle of charity is true and knowable a priori, that is, we know a priori that we should attribute beliefs and desires to people that make the most sense given their behaviors.
2. The principle of charity is true and knowable a priori only if it is true and knowable a priori that, in general, people are by and large rational, that is, that their belief-desire-behavior triples make the most sense.

3. Therefore it is true and knowable a priori that, in general, people are by and large rational.
4. Assume for reductio that it is not a necessary truth that whenever one recognizes a normative reason for action, one has a motivating reason for action.
5. Then it is possible that a person can recognize a normative reason to do A but not have a motivating reason to do A.
6. If so, then it is also possible for a person to generally lack a motivating reason to do what he recognizes he has a normative reason to do, that is, it is possible for a person to be generally irrational.
7. But if it is possible for a person to be generally irrational, then we cannot know a priori that people are in fact by and large rational.
8. This contradicts (3).
9. Therefore it is a necessary truth that whenever one has a normative reason for action, one has a motivating reason for action.

In brief, Huemer argues that internalism is true because it is necessary to explain the fact that we have a priori knowledge of the principle of charity.

Three Problems with Huemer's Argument

I begin with the most obvious problem. Huemer thinks that the principle of charity could be correct only if we were justified in believing, in general, that people are by and large rational. But crucially, he also thinks that we *could know* a priori that the principle of charity is correct only if we *could know* a priori that people are by and large rational. Perhaps he assumes that since the principle of charity is supposed to be knowable a priori, any necessary condition for it, including that people are by and large rational, must also be knowable a priori. This turns on a general principle that if something X is knowable a priori, then any necessary condition for X is also knowable a priori. This general principle is open to question, though perhaps a modified version might get the result Huemer wants.[2] Be that as it may, I am interested in the claim that it is knowable a priori that people are by and large rational. This seems plainly wrong. It is an empirical issue, not to be settled a priori, whether any being is rational in any sense. So, for example, we seek empirical evidence about the possible rationality, or the extent of the rationality, of dolphins, elephants and chimpanzees. It may be true, as some have said, that we must assume rationality if we are to treat humans as intentional systems, or some such. Now if we define "intentional system" appropriately, it might be a priori true that intentional systems must be rational. But this just moves the issue to whether humans are intentional systems. Compare this to a common issue in ethics. It might be known a priori that persons have rights, but it is an open question whether or not something is a person in the relevant sense. So, in a classic essay on abortion, Mary Anne Warren argued that we need to distinguish between the genetic (or biological) and the moral senses of "human." She declared that "It is wrong to kill innocent human beings" is self-evident only if "human" is taken in the moral sense, roughly, as being a

[2] As a reviewer for *Philosophia* pointed out, it might be that one can know a priori that one is conscious without being able to know a priori all the necessary conditions for consciousness.

full member of the moral community. But it is not self-evident, and may not even be true, if "human" is taken in the genetic or biological sense, that is, as being a member of the human species. She then develops empirical criteria for personhood (Warren 1973). In much the same way, one could argue that humans are subject to the principle of charity only if "human" is taken to mean something like "an intentional system." If "human" is used in the biological sense, as referring to members of our species, it is not self-evident and requires empirical evidence

The first problem was that whether something is rational is an empirical issue and not knowable a priori. The second problem goes a step further and identifies a particular spot in Huemer's argument that actually introduces an empirical element. This problem turns on an initial disconnect between the principle of charity and internalism because there are two notions of rationality in play. In the context of the principle of charity, rationality has to do with a person's belief's and desires making sense given his behavior. Call this "charity-rationality." In the context of internalism, rationality involves being motivated to act on what one recognizes to be reasons for action. Call this "internalism-rationality." We can rewrite the argument using the relevant senses of "rational." To get the conclusion, a transition from one sense to the other is necessary.

1. The principle of charity is true and knowable a priori, that is, we should attribute beliefs and desires to people that make the most sense given their behaviors.
2. The principle of charity is true and knowable a priori only if it is true and knowable a priori that people are generally by and large charity-rational.
3. Therefore it is true and knowable a priori that people are generally by and large charity-rational.
4. Assume for reductio that it is not a necessary truth that people are internalism-rational, that is, it is not a necessary truth that whenever one recognizes a normative reason for action, one has a motivating reason for action.
5. Then it is possible that a person can recognize a normative reason to do A but not have a motivating reason to do A.
6. If so, then it is also possible for a person to generally lack a motivating reason to do what he recognizes he has a normative reason to do, that is, to be generally internalism-irrational.
7. But if it is possible for a person to be generally internalism-irrational, then, it is possible for a person to be charity-irrational and so we cannot know a priori that people are in fact by and large charity-rational.
8. This contradicts (3).
9. Therefore, it is a necessary truth that people are internalism-rational.

The shift from internalism-irrationality to charity-irrationality takes place in (7). But how does a possible failure of internalism-rationality yield a possible failure of charity-rationality? Huemer does not explain, but I assume the idea is this. On a Humean account, belief-desire pairs are both the only reasons for action and the only motivators. So, the principle of charity, in talking about beliefs and desires as motivators, is also talking about reasons for action. Thus we move from the claim that people are possibly internalism-irrational, that is, that it is possible that they generally are not motivated by belief-desire pairs that they recognize as reasons for action, to the claim that their

behavior cannot be explained in terms of Humean motivations, that is, belief-desire pairs. This contradicts (3) and we have our reductio.

However, this does not quite work. Internalism asserts that necessarily, if one recognizes a reason for action, then one is, to some extent, motivated to act in accordance with that reason. The denial of internalism entails that it is possible to generally lack a motivating reason to do what one recognizes one has a normative reason to do. On the Humean view, belief-desire pairs are the reasons for action. But agents may not always *recognize* their belief-desire pairs to be reasons. I may spend a lot of time acting on belief-desire pairs without thinking about reasons for action at all, as when I simply grab some chocolate. So, the assumed (for reductio) failure of internalism-rationality only entails a failure of the principle of charity for belief-desire pairs one actually *recognizes* as reasons for action, and not for belief-desire pairs generally. Now suppose that many of one's belief-desire pairs are not recognized as reasons for action. Then the failure of internalism-rationality would do little to undermine charity-rationality and the reductio fails.

I do not claim that we generally fail to recognize our belief-desire pairs as reasons for action. The case that we often do recognize them as reasons for action is strengthened with the help of the distinction between an occurrent recognition and a dispositional recognition. When I grab some chocolate without any thoughts about reasons for action, I have no occurrent recognition of my reasons for action but it is likely that I have a dispositional recognition. Still, it is possible for someone not to recognize many belief-desire pairs as reasons for action. For example, young children to whom we might wish to apply the principle of charity might not be sufficiently cognitively developed to know what a reason for action is. And it might also be true that some adults regularly do not, even dispositionally, think in terms of reasons for action. How often this occurs is an empirical matter. So whether or not there is a presumption in favor of rationality turns on an empirical issue and the principle of charity cannot be known a priori. Since Huemer's argument requires not only that the principle of charity be true, but that it be knowable a priori, the argument fails.

I now turn to the third and most interesting problem with Huemer's argument. In the previous paragraphs, I questioned whether our rationality and the applicability of the principle of charity is something that can be known a priori. I now suggest, on empirical grounds, that we are to a substantial extent not rational in the sense required by Huemer's principle. I hedge with "suggest" since the empirical evidence, though reasonably strong and growing, is relatively new. There is, of course, a history of people, such as Freud, who argue that our behavior is the result of unconscious processes and hence not explainable in terms of our beliefs and desires, if these must be conscious ones. But I resort to more recent work. The sorts of problems I have in mind are similar to ones Stich has with a principle of charity for interpreting people's inferences. This principle says, roughly, that we should attribute to people inferences that are rational. His criticism turns on the existence of experimental work showing systematic irrationality in our inferences. For example, there are experiments that show people regularly violate basic tenets of reasoning about probabilities. The probability of a compound event cannot be greater than the probability of the components, but people regularly infer that it is. There are even studies that show people regularly estimate the probability of a sequence of events to be greater than the probability of the least likely

event in the sequence (Stich 1985, 118–9). I provide comparable empirical arguments against Huemer's principle of charity.

There is empirical evidence that, in many cases, the beliefs and desires that would make the most sense of behavior are not the ones people actually have. In recent years, psychologists have argued that human thinking and action is often governed by heuristics. I will not define exactly what a heuristic is and will rest content with general remarks from leading heuristics theorists. Daniel Kahneman, writes that psychologists now think that there are two modes of thinking, system 1 and system 2 (Kahneman 2011, 20–21).[3] System 1 often uses heuristics and operates automatically and quickly with little or no effort or sense of voluntary control. System 2 allocates attention to effortful mental activities such as complex computations. Examples of activities Kahneman attributes to System 1 are detecting one object is more distant than another, detecting hostility in a voice, and recognizing that a 'meek and tidy soul with a passion for detail' resembles an occupational stereotype, perhaps librarian rather than hockey player. This last is employed in experiments showing how easily System 1 thinking can lead to certain kinds of errors by using the "representativeness heuristic." Another feature of heuristics is emphasized by Gerd Gigerenzer. He speaks of heuristics as being fast and frugal. Frugality has to do with the use of limited information and processing power. For example, there is the gaze heuristic which is employed by someone catching a fly ball. We do not go through a complex mathematical calculation. The heuristic is to fix your gaze on the ball, start running, and modify your speed so that the angle of gaze is constant (Gigerenzer 2008, 7). There are other heuristics Gigerenzer mentions. There is the 'don't break ranks' heuristic which, roughly, dictates that one keep in step with what the members of one's group are doing. And there is the 'follow the default' heuristic according to which, if there are several options to chose among, one of which is the default option, one will do nothing and go with the default.

How does this undermine Huemer's principle of charity? Jonathan Haidt makes a claim specifically about moral judgment but which is generalizable. Usually, moral judgment is a matter of what he calls "intuition" by which he means a rapid, largely emotional response, say, a revulsion at incest. "The most important distinctions [between intuition and reasoning] . . . are that intuition occurs quickly, effortlessly, and automatically, such that the outcome but not the process is accessible to consciousness, whereas reasoning occurs more slowly, requires some effort, and involves at least some steps that are accessible to consciousness" (Haidt 2001, 818). The important thing here is that the intuitive responses that underlay much of our thinking are caused by processes that are not generally available to consciousness. Once intuitive judgments are made, if we are called upon to justify them, say by other people, we engage in post hoc reasoning. This reasoning is typically "lawyerly." We have made the moral judgment or performed an action for reasons not available to us and it is the job of reasoning to find something to say that sounds like a reasonable explanation that we can give to others, and sometimes to ourselves. Haidt admits that some people sometimes actually engage in reasoning to help them make up their minds about moral issues, but he thinks this is rare. The job of reasoning is mainly to find arguments to support intuitive conclusions. Haidt's view is complex, but we do not need the details. If Haidt is right that judgments and behaviors are often the results of processes that are

[3] Philosophers have long been acquainted with such distinctions through the work of R. M Hare (1981).

not available to consciousness, then accurate explanations of judgments and behaviors will be in terms of factors the agent is not aware of and, if brought to her attention, might reject.

Here is an example. Haidt describes research showing that people form impressions of others in as little as five seconds. Longer observation and deliberation rarely changes these impressions (Haidt 2001, 820). Were these judgments of politicians, and open to challenge by others, one might, after a time, come up with reasons to think that one's judgments were true, say, policies and personal qualities of the politician. But in reality, one's mind often was made up after a brief observation.[4] Haidt has himself done research on people's moral judgments, especially about certain vignettes that can arouse disgust (Haidt 2001, 823–4). He has shown that when people are asked to explain their reasons for their judgments about these cases, they are often dumbfounded and do not know what to say. This suggests that their judgments are the results of processes not consciously available to them. When asked for justifications, people often come up with explanations that could not have mattered and miss ones that might matter. This phenomenon of people, post hoc, coming up with some reason or other for their intuitive judgments points to a general phenomenon. Haidt cites Gazzaniga who claims that the brain is so good at devising post hoc explanations for our own behavior that there appears to be an interpreter module which, lacking access to the real heuristics being employed, provides constant socially appropriate hypotheses to explain why the self might have acted or judged as it did (Gazzaniga 1985). He compares this to the sort of confabulation split-brain patients show. The left hand is guided by the right brain to perform an action. The relevant information is not transmitted to the left brain because of the split, but the verbal centers in the left brain make up stories to explain what the left hand is doing, explanations that can be shown to be confabulations. If Haidt and others are right, this is just a dramatic example of a common phenomenon.

Gigerenzer provides an example that can be employed to test the principle of charity, one in which the explanation that makes the most sense of behavior is in terms of a heuristic that is not available to the agents and which they would explicitly reject (2008, 12). It is the case of the British magistrates. In England and Wales small groups of magistrates make decisions on whether to offer bail to a defendant or to make a punitive decision such as imprisonment. These magistrates are usually untrained members of the community. The bail decision has to do with the defendant's trustworthiness to turn up for court hearings, and so on, and not with guilt. The magistrates are required to pay attention to the nature and seriousness of the offense, to the character, community ties, and bail record of the defendant, and so on. Several hundred trials were observed. The magistrates took an average of 10 min to make their decisions. When asked what underlies their judgments, they typically answered that they thoroughly examined all the evidence in order to treat the defendant fairly. They sometimes spoke of the enormous weight of balancing information and of how each case is an individual case. However, an examination of their decisions indicated that the decisions were not correlated with any information that was relevant to due process, even though the magistrates had this information. Rather, the vast majority of decisions turned on a simple heuristic which Eigerenzer calls a fast and frugal tree heuristic, one that does not

[4] For a discussion of some research on the role of first impressions specifically in politics, see Christopher Olivola and Alexander Todorov (2010).

make reference to the sorts of information that the magistrates claimed to take into account. This heuristic involves a series of questions. Did the prosecution request conditional bail or oppose bail? If yes, the decision is punitive. If no, move to the next question. Did the previous court impose conditions or remand in custody? If yes, the decision is punitive. If no, move to the next question. Did police impose conditions or remand in custody? If yes, the decision is punitive. If no, the decision is not punitive and bail is granted.

This simple heuristic makes most sense of the Magistrates' decisions. One might respond that although the heuristic is *consistent* with the behavior of the Magistrates, it is more difficult to show that their decisions are actually based on it. The case that their decisions is actually based on this heuristic can be strengthened, however, if we can give a plausible explanation as to why the Magistrates would be using such a heuristic, even if they do not realize it. Gigerenzer sketches such an explanation. Briefly, he says that the institution in which the magistrates operate does not or cannot provide real feedback as to how well they are doing. The heuristic they use suggests that they therefore shift, without realizing it, to solving a different problem, protecting themselves rather than the defendant. The only time magistrates can be proven wrong is if they release a suspect who goes on to commit a crime while on bail. In such a case, the magistrates can defend themselves by arguing that neither the prosecution nor a previous court had requested a punitive decision. Gigerenzer writes "An analysis of the institution can help to understand the nature of the heuristics people use and why they believe they are doing something else" (2008, 17).

So if the principle of charity is true, the beliefs and desires of the magistrates should reflect this heuristic. But they do not. The magistrates are unaware of the heuristic they are actually following. One could respond that counterexamples like this could only refute an overly strong principle of charity. They cannot refute a more modest principle which only entails that rationality is an overridable presumption. But if this sort of phenomenon is sufficiently widespread, as the empirical evidence suggests, we are forced into a very modest principle of charity indeed, one that does not lend support to internalism. Whether there is even a justified presumption that we should interpret people's behavior in terms of beliefs and desires that make the most sense of that behavior is now open to question. At the very least, whether any principle of charity is true, and how modest it must be if it is true, turns heavily on empirical evidence. It could turn out that empirical evidence actually supports a principle of charity over a large range of cases, but this would not help Huemer. For the defense of internalism turned not just on the supposed truth of the principle of charity but crucially on the supposed fact that it could be known a priori.

I do not want to say that Huemer's principle of charity has no role in our efforts to interpret behavior. David Henderson has a plausible approach. He says that the principle of charity is subservient to, and perhaps reducible to, a principle of explicability (1987, 236). Henderson is talking about the use of charity in the creation of translation manuals. Roughly, his view is that we should begin with a principle of charity in producing a first approximation translation manual, though the principle gets more and more marginalized as we move to more refined translation manuals. The reasons we begin with a principle of charity is that such an assumption makes it likely that we will render a person's speech explicable. For example, about our translation of truth-functions for simple sentences he writes "[w]hat leads us to choose the charitable

option in translating truth-functions is the realization that a translation-*cum*-explanation for source-language speakers' behavior will be more readily obtained when following the charitable option" (Henderson 1987, 245). But even in our construction of a first approximation translation manual, we should leave significant room for error and hence for uncharitable translation. He tells us that, given our best, empirically based understanding of such things, certain kinds of errors are quite unlikely and perhaps even inexplicable. If a translation makes it appear that those errors have been committed, then that counts, perhaps conclusively, against the translation. But if errors attributable to speakers can be explicable in terms of our reasonably well confirmed theories about psychology and so on, then we may give translations that involve these mistakes (Henderson 1987, 237–9). Indeed, he mentions the efforts of Nisbett and Ross to explain errors in terms of heuristics (Nisbett and Ross 1980). So even in the early stages of constructing a translation manual, empirical evidence counts heavily on the issue of whether or not to be charitable in our translations. We should translate to maximize explicability and not to maximize charity (Henderson 1987, 227).

We should adopt the same approach for Huemer's principle of charity and subsume it under a general principle of explicability. We should interpret behavior to make it explicable, given our best understanding of things like psychology. Just as it is usually inexplicable why someone would get a simple inference from a simple conjunction wrong, so is it usually inexplicable why someone who is eating chocolate would confuse the motions of her chewing jaws with the motions of her arms swimming the English Channel. When behavior is best explained in terms of a person's beliefs and desires, then we should explain it in terms of the person's beliefs and desires. If the best explanation of the magistrates' decisions is our fast and frugal tree heuristic, then that is what we should go with. And sometimes we best explain someone's behavior in terms of serious delusions or disconnects between desires and actions.

So far I have pointed out one way that behavior can be explained uncharitably, when it is the result of heuristics that are not available to consciousness. I will mention one other way which has recently been in the news. It is speculative but highlights the empirical character of the issue. It has been known for some time that parasites can modify the behavior of insects such as ants as part of a reproductive strategy. For example, the lancet fluke causes ants to climb to the top of a blade of grass where they are more likely to be eaten by a cow or sheep, which is necessary for the fluke to complete its life cycle. Daniel Dennett made use of this example, arguing that 'memes,' especially (and rather uncharitably) religious memes, can do something similar to humans (Dennett 2007). But, more to the point here, researchers have shown that creatures besides insects can be controlled by parasites. For example, Toxoplasma gondii can only complete its life cycle if ingested by a cat. The cysts of the parasite pass into soil through the feces of an infected cat where they are ingested by rats. The parasites cause rats to lose their natural fear of cats, and even to be attracted to the smell of cat urine, making it likely that the rat will be eaten by a cat. Researchers speculate that mind control might be a common strategy parasites employ for reproduction. And what is interesting here is that the toxoplasma gondii parasite infects many creatures besides rats, including millions of humans, and may have significant effects on their behavior (Webster et al. 2013 and for a recent news report, see Zimmer 2014). This is, once again, rather speculative and no one thinks that human behavior is generally

controlled by parasites in this way. But it highlights the extent to which the domain of a principle of charity is an empirical and not an a priori matter.

My conclusion is that whether or not people are rational, in the sense of charity-rational, is an empirical matter and that the empirical evidence suggests that for some significant range of cases, we are not. But what of Huemer's explicit argument that it cannot be an empirical matter? He wrote that the only way to test our rationality (in the sense of charity-rationality) empirically is to either observe or infer beliefs and desires and to check whether they accord with behavior. But we cannot directly observe beliefs and desires. Nor can we infer them without using the principle of charity itself. Hence the application of the principle of charity is not an empirical matter. Actually, Huemer only argues that we can not know his principle of charity to be *true* empirically. He writes, "[a]t most, what could be discovered empirically is that people exhibit behavior which is *capable* of a rational interpretation, but we cannot determine empirically that interpretations of people as rational are *pro tanto* better than competing interpretations of people as irrational" (2005, 160). But presumably the brief argument is supposed to show it cannot be *refuted* empirically either. So it is not an empirical matter at all.

I have two comments about this. First, this is not something Huemer himself can consistently hold. For he thinks that there are times when we can reasonably conclude, presumably on empirical grounds, that someone is irrational in the sense that her beliefs and desires (or beliefs, desires, prudential reasons for actions, and impartial reasons for action) are out of sync with her behavior. And we can do this without direct access to her beliefs and desires or access that turns on the principle of charity since we are concluding that it does not apply to the case in hand. He does little to explain when or how we can know that someone is not charity-rational, but he could appeal to the sorts of evidence Gigerenzer appealed to in the British magistrates example. Their behavior is best explained by a fast and frugal tree heuristic, but they indignantly, and with apparent honesty, deny it. So Huemer himself is committed to there being some way to access beliefs and desires without appeal to the principle of charity.

Second, not only is it inconsistent for Huemer to press this argument, the argument itself is flawed. It turns on the notion that we cannot determine whether or not beliefs and desires can be used to explain behavior unless we either have direct access to those beliefs and desires or infer their existence using a principle of charity. But this is wrong. No doubt we spontaneously explain a great many behaviors charitably in terms of beliefs and desires, but it does not seem true that we must employ a principle of charity to explain behavior. We can, again, take a page from Henderson. What is really in play here is a general principle of explicability. We spontaneously employ explanations in terms of beliefs and desires in so far as they offer decent explanations. When other factors seem to offer better explanations we do and should appeal to them. Some people have found it natural to explain behaviors in other terms, even if we put aside behavior controlling parasites. People have, for example, often explained behavior in terms of influence by gods or by what were thought to be divine entities such as the stars and planets. For a number of years, and even today, some folk find it plausible to explain behavior in terms of struggles between ego, id and superego. Today explanations in terms of heuristics which are often not available to consciousness are common. Charity might find a role in the explanation of some behavior, as Henderson thinks it does. But whether it does, and to what extent it does, is an empirical matter. If this is right, the supposed a prioricity of the principle of charity cannot be used to support internalism.

References

Brink, D. (1989). *Moral realism and the foundations of ethics*. Cambridge: Cambridge University Press.

Dennett, D. (2007). *Breaking the spell*. London: Penguin.

Gazzaniga, M. S. (1985). *The social brain*. New York: Basic Books.

Gigerenzer, G. (2008). Moral intuition = Fast and frugal heuristics? In W. Sinnott-Armstrong (Ed.), *Moral psychology volume 2, the cognitive science of morality: Intuition and diversity* (pp. 1–26). Cambridge: MIT Press.

Haidt, J. (2001). The emotional dog and its rational tail: a social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Hare, R. M. (1963). *Freedom and reason*. New York: Oxford University Press.

Hare, R. M. (1981). *Moral thinking: Its levels, methods and point*. Oxford: Clarendon.

Harman, G. (1975). Moral relativism defended. *The Philosophical Review, 84*(1), 3–22.

Henderson, D. (1987). The principle of charity and the problem of irrationality. *Synthese, 73*, 225–252.

Huemer, M. (2005). *Ethical intuitionism*. New York: Palgrave MacMillan.

Kahneman, D. (2011). *Thinking fast, thinking slow*. New York: Narrar, Straus and Giroux.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs: Prentice Hall.

Olivola, C., & Todorov, A. (2010). Elected in 100 milliseconds: appearance-based trait inferences and voting. *Journal of Nonverbal Behavior, 34*, 83–110.

Prinz, J. (2007). *The emotional construction of morals*. Oxford: Oxford University Press.

Stevenson, C. (1937). The emotive meaning of ethical terms. *Mind, 46*, 14–31.

Stevenson, C. (1963). The nature of ethical disagreement. In *Facts and values* (pp. 1–9). New Haven: Yale University Press.

Stich, S. (1985). Could man be an irrational animal? Some notes on the epistemology of rationality. *Synthese, 64*(1), 115–135.

Thagard, P., & Nisbett, R. (1983). Rationality and charity. *Philosophy of Science, 50*(2), 250–267.

Warren, M. A. (1973). On the moral and legal status of abortion. *The Monist, 57*, 43–61.

Webster, J. M., Kaushik, G. B., & McConkey, G. (2013). Toxoplasma gondii infection, from predation to schizophrenia: can animal behavior help us understand human behaviour. *The Journal of Experimental Biology, 216*, 99–122.

Zimmer, C. (2014). Parasites practicing mind control. http://www.nytimes.com/2014/08/28/science/parasites-practicing-mind-control.html. Accessed 3 Dec 2014.